Measuring syntactical variation in Germanic texts

Wilbert Heeringa
Fryske Akademy, The Netherlands

Femke Swarte
Faculty of Arts, Applied Linguistics, University of Groningen, The Netherlands

Anja Schüppert
Faculty of Arts, European Languages and Cultures, University of Groningen, The Netherlands

Charlotte Gooskens
Faculty of Arts, Applied Linguistics, University of Groningen, The Netherlands and School of Behavioural, Cognitive and Social Sciences, University of New England, Australia

Abstract

We present two new measures of syntactic distance between languages. First, we present the 'movement measure' which measures the average number of words that has moved in sentences of one language compared to the corresponding sentences in another language. Secondly, we introduce the 'indel measure' which measures the average number of words being inserted or deleted in sentences of one language compared to the corresponding sentences in another language. The two measures were compared to the 'trigram measure' which was introduced by Nerbonne & Wiersma (2006, A Measure of Aggregate Syntactic Distance. In Nerbonne, J. and Hinrichs, E. (eds.) Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006, pp. 82–90.). We correlated the results of the three measures and found a low correlation between the results of the movement and indel measure, indicating that the two measures represent different kinds of linguistic variation. We found a high correlation between the results of the movement measure and the trigram measure. The results of all of the three measures suggest that English is syntactically a Scandinavian language. Because of our unique database design we were able to detect asymmetric relationships between the languages. All three measures suggest that asymmetric syntactical distances could be part of the explanation why native speakers of Dutch more easily understand German texts than native speakers of German understand Dutch texts (Swarte 2016).

Correspondence:
Wilbert Heeringa, Fryske Akademy, P.O. Box 54, 8900 AB. Ljouwert, The Netherlands,
E-mail:
wheeringa@fryske-akademy.nl

Digital Scholarship in the Humanities © The Author 2017. Published by Oxford University Press on behalf of EADH.

doi:10.1093/llc/fqx029

1 Introduction

Textometry is a discipline in which knowledge is derived from corpora without predefined information models. MacMurray and Leenhardt (2011) describe textometry as an approach in which 'a text possesses its own internal structure that would be difficult to analyze by manual means alone. By applying statistical and probabilistic calculations directly to the textual units of comparable texts in a corpus it becomes possible to analyze patterns and trends that would otherwise be obscured by the quantity of the textual units' (p. 606). And 'Textometry consists of seeing the document through a prism of numbers and figures, producing information on the frequency counts of words, otherwise known as occurrences, whereas forms are a single graphical unit corresponding to several instances in the text' (p. 606, see also Lebart and Salem (1994) and Tufféry (2007)).

In this article we utilize written texts for revealing language variation. Language variation at different linguistic levels become apparent to a large extent when comparing written texts in different languages, especially lexical, orthographic, and syntactical differences.

Lexical differences are differences in vocabulary or lexicon. In the following example English and German do not have any cognates, apart from the articles:

English: The boy teased the dog.
German: Der Jungen neckte den Hund.

On the other hand, pairs of sentences can be found where for each English word a German cognate is found. Cognates are words which have a common etymological origin and normally a similar shape. Example:

English: The man saw a house.
German: Der Mann sah ein Haus.

In this example differences are orthographic differences. Orthographic differences may reflect historical developments of the pronunciation, for example, English *saw* versus German *sah*. However, orthographic differences do not always reflect linguistic differences, they may also be the result of differences in spelling conventions, for example, English *house* versus German *Haus*.

Syntax is 'the study of the principles and processes by which sentences are constructed in particular languages' (Chomsky 1957, p. 11). Between Germanic languages like English and German relatively large syntactical differences can be found, for example:

English: Then she said that she will come tomorrow
German: Dann sagte sie dass sie morgen kommen wird

There exist several studies that have proposed how to measure lexical, orthographic, and syntactic distances using parallel corpora. For example, Van Bezooijen and Gooskens (2005) measured lexical distances between Dutch, Afrikaans, and Frisian on the basis of written texts. They also measured orthographic distances using the same material. Zulu, Botha, and Barnard (2008) measured orthographic distance between eleven South African languages. A procedure for measuring syntactical distances between language varieties was introduced by Nerbonne and Wiersma (2006), who provided a foundation for measuring syntactic differences between corpora. Their method uses part-of-speech (POS) trigrams as an approximation to syntactic structure. The frequencies of the trigrams of two corpora are compared for statistically significant differences.

In this article we focus on the measurement of syntactical distances between a small set of five Germanic languages. We will apply the method of Nerbonne and Wiersma (2006) and refer to this as the 'trigram measure' throughout this article. In addition, we introduce two new methods for measuring syntactical variation. Using the first method, we measure the average number of word positions that a word in a sentence in language *A* has moved compared to the corresponding sentence in language *B*. We call this the 'movement measure'. The second method measures the average number of words found in a sentence in language *A* that is missing in the corresponding sentence in language *B*, and the number of words in a sentence in language *B* that is missing in the sentence in language

A. In other words, the number of words which is *in*serted or *del*eted in a sentence in language *A* compared to the corresponding sentence in language *B* is measured. We call this the 'indel' measure.

We will compare the results of the two methods to results of the trigram method to answer the following questions:

(1) Do the movement measure and the indel measure yield different results?
(2) Does the trigram method resemble one of the other methods in particular?

We focus on the Germanic language group, more specifically on Danish, Dutch, English, German, and Swedish. In Section 2 we give a brief overview of related research concerning syntactical measurements. Section 3 describes the data source and the way in which syntactical distances are measured. The results of the distance measurements are presented in Section 4. In Section 5 the research questions are addressed. Finally, general conclusions will be drawn in Section 6. In this section we will also discuss how the methods can be validated.

2 Previous Research

To measure syntactical distances between languages we explored literature to find a suitable distance measure. We found two kinds of approaches dominating. One is based on categorical syntactical features. This approach is typically used when material from dialect atlases is used. Another is based on counting and comparing frequencies of trigrams of POS tags. This approach works well when large corpora are available with the words being tagged. The two approaches are discussed in Sections 2.1 and 2.2, respectively. In Section 2.3 we will motivate our choice.

2.1 Categorical syntactic variables

Spruit (2008) measured syntactic distances between 267 local Dutch varieties, using data from two volumes of the *Syntactic Atlas of the Dutch Dialects*. The atlas volumes contain a large number of maps showing the geographic distribution of syntactic phenomena. The maps in the first volume represent 510 binary syntactic features, and those in the second volume represent 672 syntactic features, in all 1,182 features. An example concerns the complementizer of the comparative if-clause in the Dutch sentence *Het lijkt wel alsof er iemand in de tuin staat*, 'It looks as if there is someone in the garden'. Four examples of binary features are complementizer = *of*, complementizer = *of dat*, complementizer = *dat*, complementizer = *alsof*. Each feature is either true or false, and therefore binary. The distance between two dialects was equal to the total number of shared features, and therefore, the distance will vary between 0 and 1,182.

Szmrecsanyi (2008) investigated variation in British English dialects by using the Freiburg English Dialect Corpus (FRED), a naturalistic speech corpus sampling interview material from 162 different locations in thirty-eight different counties all over the British Isles, excluding Ireland. FRED consists of 370 texts, which total about 2.5 million words of text.[1] The corpus was analysed to obtain text frequencies of sixty-two morphosyntactic features, yielding a structured database that provided a sixty-two-dimensional frequency vector per locality. The feature frequencies were subsequently normalized to frequency per 10,000 words (because textual coverage in FRED varies across localities) and log-transformed to deemphasize large frequency differentials and to alleviate the effect of frequency outliers. The resulting 38×62 table (on the county level—that is, thirty-eight counties characterized by sixty-two feature frequencies each for the full data set) was converted into a 38×38 distance matrix using Euclidean distance—the square root of the sum of all squared frequency differentials—as an interval measure. This distance matrix was subjected to cluster analysis to find dialect groups.

Grieve (2016) analysed a word corpus representing the letter to the editor register as written between 2000 and 2013 in 240 cities from across the USA. The letters were downloaded from the online archives of one or more newspapers published in 240 cities. A total of 135 grammatical alternation variables were measured and mapped across the 240 city sub-corpora. An alternation variable is 'a set of distinct linguistic forms that have the same referential meaning' (p. 36). The percentage of each variant is calculated as the quotient of the total

number of tokens of that variant in the corpus and the total number of tokens of all the variants of that alternation variable in the corpus, multiplied by 100 (see also Grieve 2009).

Spruit (2008), Szmrecsanyi (2008), and Grieve (2009, 2016) used syntactic alternation variables (or linguistic variables) which were found in a dialect atlas (Spruit, 2008) or derived from written text corpora (Szmrecsanyi, 2008; Grieve, 2009, 2016).

2.2 Frequencies of POS categories

Hirst and Feiguina (2007) presented a method for authorship discrimination that is based on the frequency of bigrams of syntactic labels that arise from partial parsing of the text. With this method the authors obtained a high accuracy on discrimination of the work of Anne and Charlotte Brontë (Brontë, 1847, 1848, 1853), both alone and combined with other classification features. High accuracies are achieved even on fragments of short texts of little more than 200 words long.

While Hirst and Feiguina (2007) focussed on determining the authorship of texts, Nerbonne and Wiersma (2006), Lauttamus *et al.* (2007), Wiersma *et al.* (2010), and Nerbonne *et al.* (2010) measured the impact of L1 on L2 syntax in second language acquisition on the basis of corpora of English of Finnish Australians. They presented an application of a technique from language technology to tag a corpus automatically and to detect syntactic differences between two varieties of Finnish Australian English, one spoken by the first generation and the other by the second generation. The technique compares frequencies of trigrams of POS categories as indicators of syntactic distance between the varieties and then examine potential effects of language contact. The frequency vectors were compared and analysed by using a permutation test, which resulted in both a general measure of difference and a list with the *n*-grams that are most responsible for the difference. The findings showed syntactic 'contamination' from Finnish in the English of the adult first-generation speakers of Finnish ethnic origin. The results show that we can attribute some interlanguage features in the first generation to Finnish substratum transfer.

Sanders (2007) extended the method and its application. He extended the method by using leaf-path ancestors of Sampson (2000) instead of trigrams, which captures internal syntactic structure—every leaf in a parse tree records the path back to the root. The corpus used for testing is the International Corpus of English, Great Britain (Nelson *et al.*, 2002), which contains syntactically annotated speech of Great Britain. The speakers were grouped into geographical regions based on place of birth. Sanders showed that dialectal variation in eleven British regions from the International Corpus of English, Great Britain (ICE-GB) is detectable by the algorithm, using both leaf-ancestor paths and trigrams.

2.3 Our approach

Spruit (2008), Szmrecsanyi (2008), and Grieve (2016) quantified syntactical language variation by using alternation variables. When using corpora as in the case of Szmrecsanyi (2008) and Grieve (2016), a set or features need to be chosen. The choice of features may partly depend on the data, but will easily be subjective.

Given the fact that we use corpora (see Section 3.1) we prefer not to choose a set of features, but simply measure syntactical distances in terms of differences of sentence structure, regardless what features are represented by those differences. We will introduce two new measures. The first one measures the average number of word positions that a word in a sentence in language *A* has moved compared to the corresponding sentence in language *B*. The second measure measures the number of words which is *in*serted or *del*eted in a sentence in language *A* compared to the corresponding sentence in language *B*.

The methodology of Hirst and Feiguina (2007) and Nerbonne and Wiersma (2006) likewise does not require the choice of a feature set and excels in simplicity. We will also consider their methodology and compare the results of our measures with their trigram measure.

Nerbonne and Wiersma's (2006) method is sensitive only to sequential order, not to insertions, deletions, or phrase structure. Sanders (2007) clearly increased the sensitivity of the measure he

developed a great deal with respect to phrase structure. It might be argued that the movement and indel measures are potentially sensitive to higher levels of syntactic structure, perhaps even transformational structure (Chomsky, 1957).

3. Data Source and Measurement Techniques

The data used in this article were collected in the context of a research programme which aims at finding linguistic and non-linguistic determinants of mutual intelligibility within the Germanic, Romance, and Slavic language families. Within this research programme, web-based intelligibility tests were performed and linguistic distances between the languages were measured (Golubović, 2016; Swarte, 2016).

3.1 Data source

The basis of our analyses is a set of four English texts at the B1/B2 level according to the Common European Framework of Reference for Languages.[2] The texts were used as preparation exercises for the Preliminary English Test. The diploma is offered by University of Cambridge ESOL Examinations in England. The texts we use are obtained at `englishaula.com`.

The texts are translated in each of the other four languages (Dutch, Danish, German, Swedish) by native speakers of those languages. The translations are subsequently corrected by two other native speakers. All of the native speakers had completed a university education or were still studying at a university. They were aged between 20 and 40 years.

Just as the English text the four texts consist of sixty-six sentences (approximately 800 words) in total. Given five languages, we will analyse $(5 \times 4)/2 = 10$ language pairs. Our initial thought was to calculate the syntactic distance of a language pair by directly comparing the texts of the two languages to each other. However, by doing this, we would introduce a lot of noise in our data. We will illustrate this by an example. In the text *Child Athletes* we find the following sentences in English and German:

> English: Some doctors agree that young muscles may be damaged by training before they are properly developed.
> German: Einige Ärzte behaupten, dass junge Muskeln die noch nicht ausreichend entwickelt sind während des Trainings beschädigt werden können.

The two sentences have about the same meaning, but syntactically they strongly differ. However, given the English sentence, it is possible to get a more literal German translation:

> English: Some doctors agree that young muscles may be damaged by training before they are properly developed.
> German: Einige Ärzte denken, dass junge Muskeln durch Training geschädigt werden können bevor sie ausreichend entwickelt sind.

On the other hand, given the German sentence, a more literal translation in English is possible:

> German: Einige Ärzte behaupten, dass junge Muskeln die noch nicht ausreichend entwickelt sind während des Trainings beschädigt werden können.
> English: Some doctors claim that young muscles which are still not properly developed can be damaged during the training.

Since we want to model intelligibility (see Section 6), we should not calculate syntactic distances which are unnecessarily large. A reader who reads a sentence in a closely related language, will likely try to match the sentence with the most literal translation in his/her own language.

Therefore, to obtain the data set that our analysis will be based on, each of the available texts in Danish, Dutch, English, German, and Swedish are 'translated back' in each of the other languages as literally as possible. Importantly, the texts are translated as literally as possible with respect to syntax, but not necessarily with respect to lexicon, as this is not within the scope of this article. However, the translations are made so that the sentences are still grammatically correct. These translations are language specific, i.e. the Danish text is translated in a different way from Swedish than from German, for example. Note that we modified only the targets

Table 1 Each of the available texts in Danish, Dutch, English, German, and Swedish are translated in each of the other languages as literally as possible

Text in:	**Is translated as literally as possible in:**			
Danish	Dutch	English	German	Swedish
Dutch	Danish	English	German	Swedish
English	Danish	Dutch	German	Swedish
German	Danish	Dutch	English	Swedish
Swedish	Danish	Dutch	English	German

since what a potential reader would encounter ought not to be changed. In this way we obtain twenty-five corpora, each corpus consists of sixty-six sentences. This is schematically shown in Table 1.

When calculating syntactic distances and modelling for example a German reader who is reading a Danish text, we calculate the distance between the original Danish text and the 'literal' German translation of this text. Likewise, when modelling a Danish reader who is reading a German text, we calculate the distance between the original German text and the 'literal' Danish translation of this text. By using this method, we ensure that we model syntactic intelligibility in listeners that are confronted with a closely related language, instead of reporting syntactic distances between two sometimes arbitrarily formulated translations. This example also shows that our data set enables us to find asymmetries in syntactic relationships: reading Danish by a German may be easier or more difficult than reading German by a Dane from a syntactic point of view.

We manually tagged the words in each of the corpora for syntactic word class. We used the tags which are listed in Table 2. In the corpus the first word of a sentence is preceded by a $, and the last word of a sentence is followed by a #. Thus we marked the beginning and the end of a sentence, which will especially play a role in the trigram measure (see Section 3.2.3).

Table 2 When tagging the words in the corpus we distinguished thirteen word classes

Tag	**Word class**	**Examples**
$		Beginning of sentence
noun	Noun	*parents, trouble, sport*
verb	Verb	*are, allow, become*
mod	Modal verb	*can, could, will, be*
adj	Adjective	*young* children
adv	Adverb	starting *young*
pron	Pronoun	*it, they, you, my, that*
prep	Preposition	*for, at, of*
conj	Conjunction	*and, also, but, if, or*
num	Numeral	*five, most, all*
det	Determiner	*a, the, another, this*
int	Interjection	*hello, please*
to	*To* before infinitive	*to* prevent, *to* do
abbr	Abbreviation	*etc.*
#		Ending of sentence

The beginning and ending of a sentence are also marked.

3.2 Measuring syntactical distances

3.2.1 *The movement measure*

When reading texts in an unknown closely related language, the reader may find that words have 'moved', i.e. occur at a different position in the sentence than expected by the reader. For example:

> **Dutch text:** Wanneer geen hulp gegeven kan worden ...
> **German reader:** Wenn keine Hilfe gegeben werden kann ...

Dutch *worden* corresponds with German *werden*. A native speaker of German expects this word between *gegeven* and *kan*, but it has 'moved' to the position following on *kan*. We find that *werden* in the German sentence has moved two positions forward in the Dutch sentence where *werden* is translated as *worden*.

The movement measure measures the average number of word positions that a word in a sentence in language *A* has moved compared to the corresponding sentence in language *B*. When comparing

languages *A* and *B*, we have to compare the sixty-six sentences in the corpus of language *A* with the sixty-six sentences in the corpus of language *B*. The procedure consists of four steps.

3.2.1.1 *First step: Code word correspondences*

We add codings to words so that words in a sentence in language *A* which correspond with words in the corresponding sentence in language *B* have the same codes. Assume German *Nach einiger Zeit wird es einfacher werden* and English *After some time it will become easier*. We would add codings as follows:

> **German:** 1-Nach 2-einiger 3-Zeit 4-wird 5-es 6-einfacher 7-werden
> **English:** 1-After 2-some 3-time 5-it 4-will 7-become 6-easier

The codings make clear that for example German *wird* corresponds with English *become*, and German *einfacher* corresponds with English 'easier', even if the German and English words are found at different positions and are no cognates of each other.

3.2.1.2 *Second step: Align the sentences*

With Levenshtein distance each of the sixty-six sentences in the corpus of language *A* is aligned to a corresponding sentence in the corpus of language *B*. The Levenshtein distance is a numerical value defined as the cost of the least expensive set of insertions, deletions, and substitutions needed to transform one string into another (Kruskal, 1999). Assume the sequences *abc* and *ecd*, the strings are aligned as follows:

a	b	c	
	e	c	d
del.	sub.	match	ins.

We find that *abc* can be changed into *ecd* by a deletion, a substitution, and an insertion.

When comparing the sentences, the Levenshtein algorithm will align the sentences so that words which have different lexical codings will not be aligned to each other. Words which have the same lexical codings but belong to different word classes, may be aligned to each other, but this gives a distance score of 0.5. For example, when English *time of year* is compared to German *Zeit des Jahres*, English *of* is matched with German *des*. Since *of* is a preposition and *des* a determiner, a distance score of 0.5 is given. This does not happen frequently. The insertion or deletion of a word gives a distance score of 1.0. Returning to our German/English example, Levenshtein will align the sentences as follows:

	1	2	3	4	5	6	7	8	9
German:	1-Nach	2-einiger	3-Zeit	4-wird	5-es			6-einfacher	7-werden
English:	1-After	2-some	3-time		5-it	4-will	7-become	6-easier	
	match	match	match	del.	match	ins.	ins.	match	del.

3.2.1.3 *Third step: Quantify the movements*

When considering the alignment, we find that German *wird* has moved from position 4 to position 6, which is a movement of 6–4 is two positions. German *werden* has moved from position 9 to position 7, which is a movement of 9–7 is two positions. The total number of movements is $2+2$ is four positions.

In the measurement described above, we assume that the further a word is moved the more it will affect intelligibility negatively. We consider two alternatives where larger movements are weighted relatively less heavily, namely, logarithmic and binary distances. When calculating logarithmic distances, we calculate the natural logarithm. In our case the distance becomes $ln(4) = 1.386$. The effect is that large movements count relatively less strongly than small distances. With binary we mean that we determine whether a word has moved or not. We simply count the number of words that has moved to another position in a sentence.

3.2.1.4 *Fourth step: Calculate the aggregate*

When comparing two languages, we calculate the number of movements for sixty-six sentence pairs. The aggregated distance is the average of the sixty-six movement measurements.

3.2.2 *The indel measure*

When reading a text in a closely related language, a reader may find that words are added or removed in

comparison to the closest possible sentences which the reader would have used him/herself. For example:

> **German text:** Das passiert sogar nach vielen Jahren Training
> **English reader:** It happens even after many years of training

A native speaker of English expects a German equivalent of the preposition *of*, but it is lacking in the German sentence. The indel measure measures the number of word *in*sertions and word *dele*tions (word indels). The procedure consists of the same four steps as in the movement measure (see Section 3.2.1). Only the third step is different. Instead of quantifying the number of movements, the number of indels is counted. Assume the following pair of sentences being aligned by Levenshtein distance:

	1	2	3	4	5	6	7	8	9	10	11	12	13
English	You	are	not	carrying	the	weight	of	your	body		on	your	feet
German	Du			trägst	das	Gewicht		deines	Körpers	nicht	auf	deinen	Füßen

It may look as if we find three deletions (at positions 2, 3, and 7) and one insertion (at position 10). However, the English word *not* at position 3 has moved to position 10. The number of movements is $10 - 3 = 7$. We do not find any insertions, but we find deletions at positions 2 and 7. Therefore the number of indels equals 2.

3.2.3 *The trigram measure*

Nerbonne and Wiersma (2006) used a technique which utilizes frequency profiles of trigrams of POS categories as indicators of syntactic distance between the varieties. We described the method in Section 2.2 and refer to it as the trigram measure.

We illustrate the finding of trigram frequencies on the basis of a small corpus of English, which consists of two sentences: 'It would be difficult to cycle' and 'After a while it will become easier'. For the first sentence, we find six trigrams:

	It	*would*	*be*	*difficult*	*to*	*cycle*	
$	Pron	mod					
	Pron	mod	verb				
		mod	verb	adv			
			verb	adv	to		
				adv	to	verb	
					to	verb	#

For the second sentence, we find seven trigrams:

	After	*a*	*while*	*it*	*will*	*become*	*easier*	
$	Prep	det						
	Prep	det	noun					
		det	noun	pron				
			noun	pron	mod			
				pron	mod	verb		
					mod	verb	adv	
						verb	adv	#

Note that also the $ which marks the beginning of a sentence, and the # which marks the ending a sentence, can be parts of a trigram.

When we make an inventory of the trigrams found in the two sentences, we find eleven different types of trigrams:

Trigram			**Sentence 1**	**Sentence 2**	**Frequency**
$	pron	mod	x		1
pron	mod	verb	x	x	2
mod	verb	adv	x	x	2
verb	adv	to	x		1
adv	to	verb	x		1
to	verb	#	x		1
$	prep	det		x	1
prep	det	noun		x	1
det	noun	pron		x	1
noun	pron	mod		x	1
verb	adv	#		x	1

The last column is a frequency vector of trigrams of POS tags. It shows that most trigrams occur once, but the trigrams pron/mod/verb and mod/verb/adv appear twice.

In Section 3.1 we explained that we created twenty-five corpora. Now we create a frequency vector for each of the corpora. For each corpus we consider sixty-six sentences. Given thirteen classes, a marker for the beginning of a sentence and a marker for the ending of a sentence, the number of possible trigrams is $(13+1+1)^3 = 3{,}375$. Therefore a frequency vector consists of 3,375 frequencies. For trigrams which do not occur the frequency is equal to 0.

Once we have a frequency vector for each corpus, distances can be calculated for each language pair. When we want to model a Dane who is reading German, we create a frequency vector on the basis of the 'original' German corpus and a frequency vector on the basis of the 'literal' Danish translation of the German texts. The distance is calculated as 1 minus the Pearson's correlation of the two vectors. The significance of the correlation is found by means of a Mantel test.

Nerbonne and Wiersma (2006) compared and analysed frequency vectors by using a permutation test, which results in both a general measure of difference and a list with the *n*-grams that are most responsible for the difference. Our approach is simpler, but the use of a permutation test may be a topic for future work.

The trigram measure has some advantages over the two measures which we discussed above. While both the 'movement measure' and the 'indel measure' require the aligning of sentences using a procedure which needs to know which word in the stimulus language corresponds to which word in the subject language, this is not required by the trigram measure. Parallel corpora are not even required when the samples are sufficiently large.

4. Syntactic Distances between Languages

4.1 Consistency

Cronbach's α is a popular method to measure consistency or reliability. Cronbach (1951) proposed the coefficient as a lower bound to the reliability coefficient in classical test theory. The value of α indicates the extent to which a given set of items measures the same concept. Cronbach's α measures how closely related a set of items are as a group. Its values range between zero and one. Higher values indicate more reliability. As a rule of thumb, values higher than 0.7 are considered sufficient to obtain consistent results in social sciences (Nunnally, 1978).

Table 3 Cronbach's α values for four different measures obtained on the basis of five languages and sixty-six sentences per language

Measure	Cronbach's α
Movement, linear	0.966
Movement, logarithmic	0.970
Movement, binary	0.971
Indel	0.925

Our movement and indel measurements are based on five languages. When using the movement and indel measures, sentences are distinguished. In our data, we have sixty-six sentences. These are the test items. To determine whether sixty-six sentences are sufficient, we calculated Cronbach's α values. The results are shown in Table 3. All of them are high and show that a data set of sixty-six sentences is sufficiently large.

We did not calculate Cronbach's α values for the trigram measure. The trigram measure does not distinguish items (in our case: sentences), but considers the set of sixty-six sentences as a corpus, on the basis of which frequencies of trigrams are found.

4.2 Distances

In this section we show and analyse distances measured with the movement measure, the indel measure, and the trigram measure. As to the movement measure, we show results for the linear version only. Results obtained on the basis of the logarithmic version and the binary version are nearly identical; we come back to this in Section 5.

The movement distances, indel distances, and trigram distances between the five languages are given in Tables A1, A2, and A3, respectively. The distances are visualized by means of cluster analysis (Section 4.2.1) and multidimensional scaling (MDS) (Section 4.2.2).

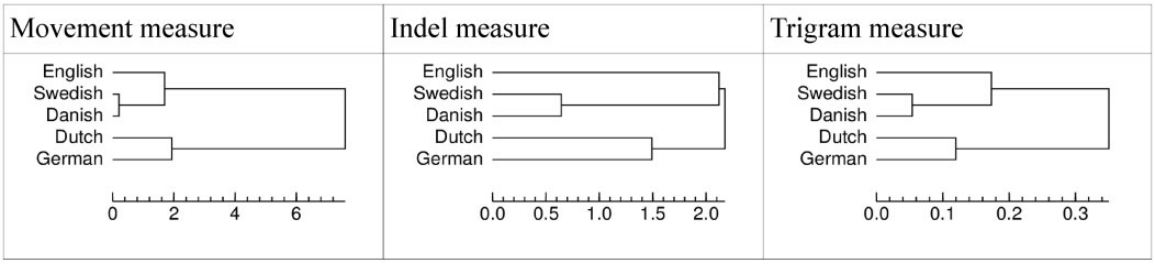

Fig. 1 Dendrograms obtained on the basis of movement distances, indel distances, and trigram distances. The tree structures explain respectively 96.2%, 83.9%, and 97.4% of the variance in the original distances

4.2.1 *Cluster analysis*

We applied hierarchical cluster analysis to the movement distances, the indel distances, and the trigram distances. The two measurements per language pair are averaged. For each measure we obtain a binary tree structure in which the varieties are the leaves and the branches reflect the distances between the leaves, known as a dendrogram (Jain and Dubes, 1988).

Several alternatives exist. We used the Unweighted Pair Group Method using Arithmetic (UPGMA) averages, since dendrograms generated by this method reflect distances which correlate most strongly with the original distances ($r = 0.96$ for movement distances, $r = 0.84$ for indel distances, and $r = 0.97$ for trigram distances), see Sokal and Rohlf (1962).

The dendrograms are shown in Fig. 1. All of them show a division between a northern group (English, Swedish, and Danish) and a southern group (Dutch, German). But in the dendrogram obtained on the basis of the indel distance the distance between English and the Scandinavian subcluster (containing Swedish and Danish) is relatively large. In the two other dendrograms English is much tighter clustered with the Scandinavian subcluster.

4.2.2 *Multidimensional scaling*

On the basis of geographic coordinates, the distances between locations can be determined. The reverse is also possible: on the basis of the known distances, an optimal coordinate system can be determined with the coordinates of the locations in it. The latter is realized by a technique known as MDS. The purpose of MDS is to provide a visual representation of the pattern of distances among a set of elements. On the basis of distances between a set of elements a set of points is returned so that the distances between the points are approximately equal to the original distances. The result is that on the plot, like concepts are plotted nearby and unlike concepts are distant.

Torgerson (1952) proposed the first MDS method which is known as 'classical multidimensional scaling'. The method is also described in Torgerson (1958) and is a metric procedure. 'Sammon's mapping' (Sammon, 1969) is closely related to classical multidimensional scaling. This method also tries to optimize a cost function that describes how well the pairwise distances in a data set are preserved. But Sammon's mapping is especially useful when the preservation of small distances needs to be emphasized.

We applied Sammon's MDS. Just as for cluster analysis, the two measurements per language pair are averaged. We found that MDS plots made with Sammon's MDS most closely agree with the dendrograms in Fig. 1. The plots are shown in Fig. 2. In case of the movement and trigram measure, we find a distinction between a northern group (English, Swedish, and Danish) and a southern group (Dutch and German). In case of indel measure, the situation is less clear. As to English, it cannot clearly be concluded whether this language belongs to the northern or the southern group.

4.3 Is English a Scandinavian language?

In the previous sections we find English grouped together with Danish and Swedish, especially in

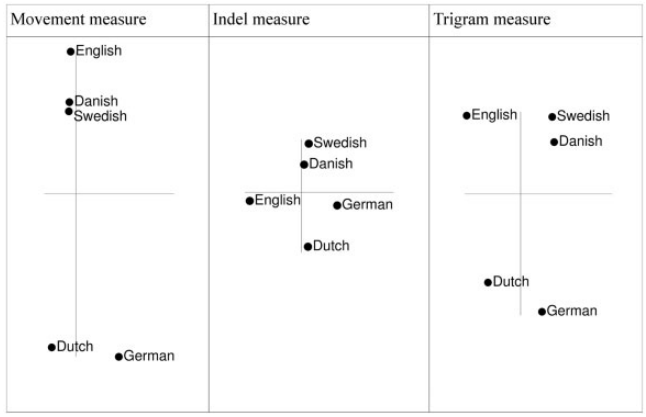

Fig. 2 MDS plots obtained on the basis of movement distances, indel distances, and trigram distances. The plots explain respectively 99.3%, 98.9%, and 99.1% of the variance in the original distances

the results of the movement measure and the trigram measure. This is remarkable, since English is usually classified as a West-Germanic language. It is assumed that English originated from the fusion of closely related dialects, now collectively termed Old English, spoken by Germanic settlers, and ultimately from their ancestral region of Angeln, a small area in the northeast of the German state of the Schleswig-Holstein and bounded on the north by German/Danish state border. The language was influenced by the Old Norse language because of Viking invasions in the 8th and 9th centuries (Baugh and Cable 1978). The large number of words having a Scandinavian origin in English is therefore attributed to language contact and heavy borrowing of Scandinavian words into Old or Middle English. Borrowing lexical words is common in contact situations.

However, Emonds and Faarlund (2014) point to the fact that many grammatical words and morphemes were also borrowed, which is unusual. Even more problematic is the fact that Middle English and Modern English syntax is of a Scandinavian rather than a West-Germanic type. Therefore, they argue that the linguistic ancestor of Middle English (and therefore Modern English) is North Germanic, with large borrowings from the Old English lexicon, rather than the other way around. According to the authors, Middle English in fact descended from the language of Scandinavians who settled in the British Isles in the course of many centuries, before the French-speaking Normans conquered the country in 1066. The vocabularies of Old English and Scandinavian fused in the 12th century during the full impact of the Norman Conquest.

The authors found evidence for this by reviewing '20 syntactic constructions where Middle English and consequently, in most cases, Modern English clearly exhibit the North Germanic patterns, showing that *English syntax is uniformly North Germanic*' (p. 131). For example, both in English and Scandinavian the object is placed after the verb:

English I have read the book
Danish Jeg har læst bogen
Norwegian Jeg har lest boken
Swedish Jag har läst boken

German and Dutch (and Old English) put the verb at the end:

German Ich habe das Buch gelesen
Dutch Ik heb het boek gelezen

Also when a to + infinitive structure is used, the object or adjective is placed after the verb in English and Scandinavian[3]:

English Rule Number Two is to pay attention
Danish Regel nummer to er at være opmærksom

Swedish Regel nummer två är att vara uppmärksam

In Dutch and German the adjective precedes the to + infinitive structure:

Dutch Regel nummer twee is om alert te zijn
German Regel Nummer zwei ist aufmerksam zu sein

English and Scandinavian can have a preposition at the end of the sentence:

English This we have talked about
Danish Det har vi talt om
Norwegian Dette har vi snakket om
Swedish Detta har vi talat om

In German and Dutch the preposition is combined with the demonstrative adverbs:

German Darüber haben wir gesprochen
Dutch Daarover hebben we gesproken

The question may rise whether there are any syntactic features in Middle English more reminiscent of Old English rather than of Scandinavian. The authors write, '... the answer is negative. The (extensive) syntactic evidence *all goes one way*' (p. 131). And they continue:

> 'Therefore, by the criterion of syntactic descent (11), Middle and Modern English are indisputably North Germanic. The "family trees" indicating that they are West Germanic, found in even the most recent sources, e.g., Miller (2012, p. 3), are all incorrect.' (p. 131)

Our results, as shown in Figs 1 and 2, confirm that English is rather a North-Germanic than a West-Germanic language when considering language variation at the syntactical level. However, as a reaction to the findings of Emonds and Faarlund (2014) a number of authors (Van Gelderen, 2016; Van Kemenade, 2016; Kortmann, 2016; Trudgill, 2016) presented arguments against this conclusion and the classification of English therefore is still an open issue.

4.4 Asymmetries

In Section 3.1 we emphasized that we do not want to calculate syntactic distances which are unnecessarily large, given our aim to model intelligibility. Given a text for language *A*, we translated this into language *B* as literally as possible, and given a text for language *B*, we translated this into language *A* as literally as possible. Next, we calculated distances *A*/*B* (representing a native speaker of language *B* who reads language *A*) and *B*/*A* (representing a native speaker of language *A* who reads language *B*).

The visualization techniques used in Sections 4.2.1 and 4.2.2 are not able to deal with two distances per language pair, i.e. distances *A*/*B* and *B*/*A*. Therefore the two measurements per language pair were averaged before applying the visualization techniques. However, that the distance tables in Appendix A show distances *A*/*B* and *B*/*A* may be different. For example, the distance of a Dutchman reading English is 8.36, and for an Englishman reading Dutch the distance is 9.79 when looking at the results of the movement measure (Table A1). Asymmetries occur when one language allows more syntactical variants than another. For example, the English sentence:

The house which he has seen

can be translated into Dutch without changing the word order:

Het huis dat hij heeft gezien

However, the Dutch sentence:

Het huis dat hij gezien heeft

cannot be translated into English without changing the word order in the final verb cluster. This is an example of asymmetry.

In this section we will show that asymmetric relationships can be found for movement, indel, and trigram measures. In Sections 3.2.1 and 3.2.2 we discussed the movement measure and the indel measure. When comparing two languages on the basis of sixty-six sentences using the movement measure, for each sentence the number of movements is counted.[4] When using the indel measure, for each of the sixty-six sentences the number of indels is counted. For both the movement and the indel measure the aggregate distance is the average of the sixty-six counts.

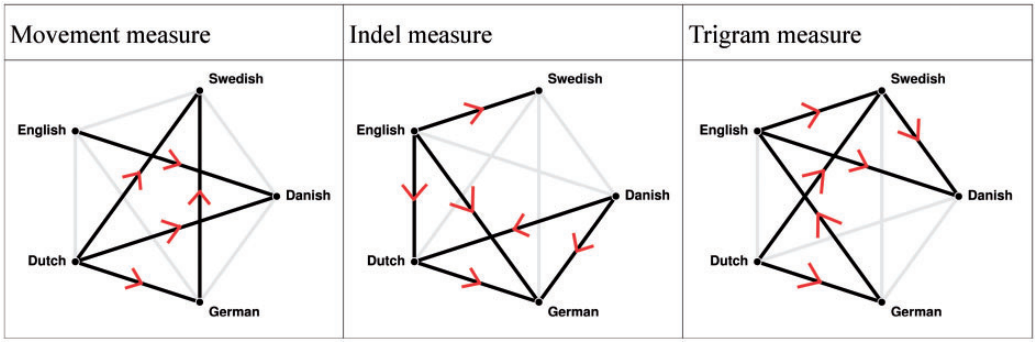

Fig. 3 Asymmetries in movement, indel, and trigram distances. An arrow from language *A* to language *B* predicts that the native speakers of language *A* significantly more easily understand language *B*, than native speakers of language *B* understand language *A*. Asymmetries are significant at least at $\alpha = 0.05$

When comparing languages *A* and *B*, we distinguish between pair *A*/*B* and *B*/*A*. When the aggregate distance *B*/*A* is smaller than distance *A*/*B*, this suggests that a native speaker of language *A* more easily understands language *B*, than a native speaker of language *B* understands language *A*. To test whether a significant asymmetric relationship exists, we need to test statistically whether the sixty-six sentence counts of *B*/*A* are smaller than the sixty-six sentence counts of *A*/*B*. We did this for each language pair by means of a paired-samples *t*-test. The results are shown in Fig. 3.

The results of the movement measure suggest that that native speakers of Dutch will more easily understand texts written in Swedish, Danish, and German than the other way around. For Danes it is predicted to be harder to read English and Dutch, and for Swedes it is harder to read Dutch and German than the other way around.

Looking at the results for the indel measure, we expect that native speakers of English more easily understand texts written in Swedish, German, and Dutch than the other way around, and Danes more easily understand texts written in Dutch and German than the opposite. For the Dutch it is easier to read German texts than for Germans to read Dutch texts.

For finding asymmetries in the trigram measurements we compared the correlations between the trigram frequency vectors (see Section 3.2.3). The inverse correlation coefficients (i.e. 1 minus the correlation) are given in Table A3, but we detect asymmetries by comparing the original correlation coefficients. When the correlation *B*/*A* is larger than the correlation *A*/*B*, this suggests that a native speaker of language *A* more easily understands language *B*, than a native speaker of language *B* understands language *A*. When comparing two correlation coefficients, we take into account that frequency vectors consists of 3,375 frequencies (see Section 3.2.3).[5]

The asymmetries are shown in Fig. 3 again. They suggest that it is easier for native speakers of Dutch to understand texts written in Swedish and German, than the other way around. For native speakers of English it may be easier to understand texts written in Swedish and Danish, than the other way around. Swedes may more easily read Danish texts, than Danes read Swedish texts. Germans may more easily understand English texts, than the English understand German texts.

The movement measure and the trigram measure share three asymmetries going in the same direction (English to Danish, Dutch to Swedish, Dutch to German). The indel measure and the trigram measure share two asymmetries (English to Swedish, Dutch to German). The movement and the indel measure share just one asymmetric relationship going in the same direction (Dutch to German). This relationship is shared by all of the three measures and could be part of the explanation why native speakers of Dutch more easily understand German

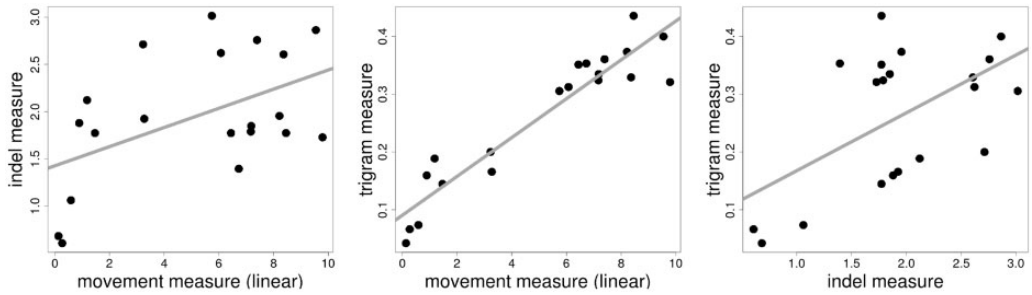

Fig. 4 Scatterplots showing the correlations between movement measures, indel measures, and trigram measures

Table 4 Correlations between five syntactical distance measures

Measure		Movement			Indels	Trigrams
		Linear	Logarithmic	Binary		
Movement	Linear		0.9959***	0.9890***	0.5008*	0.9408***
	Logarithmic			0.9979***	0.4969*	0.9605***
	Binary				0.4998*	0.9688***
Indels						0.5679**
Trigrams						

The asterisks show the significance of the correlations: * means $P < 0.01$, ** means $P < 0.001$, *** means $P < 0.0001$.

texts than native speakers of German understand Dutch texts (Swarte, 2016).

5 The Three Syntactic Measures Compared to Each Other

The results in Section 4.2 suggest that the movement and trigram measures yield similar results, and that the results of the indel measure are different. In our research question we ask how well the three measures of syntactic distance are related to each other. Therefore, we correlate the results of the measures to each other. We have five measures: the linear, logarithmic and binary movement measure (Section 3.2.1), the indels measure (Section 3.2.2), and the trigram measure (Section 3.2.3). Given five languages, for each measure we obtain a matrix which contains $5 \times 5 = 25$ distances. Since a distance of a language to itself is 0, we leave them out (see the matrices in Appendix) and consider twenty distances.

Two measures are compared to each other by correlating the results they produced. We correlated by means of the Pearson's correlation coefficient, and got p values by means of the Mantel test (Mantel, 1967). The Mantel test calculates the significance levels of correlation coefficients between distance tables while taking into account the structured, interdependent nature of distance matrices. The null hypothesis in this asymptotic test states that there is no correlation between the distances in the two matrices. The results are presented in Table 4. In this table the three kinds of movement measures (linear, logarithm, and binary) are included. The scatterplots in Fig. 4 show the correlations between the linear movement measure, the indel measure, and the trigram measure.

The table shows that the correlations between the movement measures are high, varying between 0.9890 and 0.9959. The movement measures also correlate strongly with the trigram measure. The correlations vary between 0.9408 and 0.9688. The indel measure correlates less strongly with the movement measures

(correlations between 0.4998 and 0.5008) and with the trigram measure ($r = 0.5679$).

6 Conclusions and Discussion

In Section 5 we found high and significant correlations between movement distances and trigram distances, but found lower correlations between the indel distances and the trigram distances, and between the movement distances and the indel distances.

The high correlation between the movement measures and the trigram measures is remarkable. Intuitively the movement measure may be preferred, since it precisely measures the number of positions that words have moved in a sentence. The trigram measure only considered frequencies of POS trigrams. The use of the trigram measure is much more practical, words only need to be tagged. The movement measure requires parallel corpora where words are coded so that the algorithm knows which words correspond to each other. This coding has to be done manually and is time-consuming (2–3 h for the sixty-six sentences per language pair) and not required by the trigram measure. Additionally, the trigram measure does not require parallel corpora. It only needs POS tagging, which can be automatized by using tools like 'TreeTagger' which can be applied to corpora of a large number of different languages.

We found that all of the movement measures strongly correlate with the trigram measures. On the one hand, the linear movement measure measures the exact number of positions that words have moved in a sentence. On the other hand, the binary measure simply counts the number of words that has moved to another position in a sentence. The logarithmic movement measure can be considered as an intermediate form. The fact that all of the movement measures strongly correlate with the trigram measures, may indicate that an exact counting of the number of words positions is not necessary.

We have to consider that the indel measure highly depends on the way sentences are translated. Even when translating as literally as possible (see Section 3.1), consistency among the translators of the five languages cannot be guaranteed. Therefore the indel distances are more easily flawed than the other measures.

For all of the measures it can be tested whether the relationships between languages are asymmetric. The different results found for different measures show the different nature of the measures, but most overlap was found between the movement and the trigram measure.

The initial purpose of our syntactical distance measurements was to develop a statistical model of mutual intelligibility between closely related languages. Swarte (2016) measured mutual intelligibility between five Germanic languages by means of a spoken and a written cloze test. She correlated the intelligibility scores with non-linguistic factors (attitude and exposure to the test language) and with linguistic distance measurements (lexical, phonetic, orthographic, and syntactic). The movement measure and the trigram measure correlated significantly with written and spoken intelligibility at the 0.05 level. In a stepwise regression model with all linguistic and non-linguistic factors entered, the movement measure was the only syntactic measurement that was a significant predictor of written intelligibility. No syntactic measurements were entered in the model of spoken intelligibility. However, if removing the most important predictor, exposure, from the model both the movement and the trigram measurements were included in the model. These results therefore do not help us to make a choice between these two measures, they rather seem to be complementary to each other. In future work we will look further into the relationship between syntactic distances and intelligibility. For example, it is important to include other language families to validate the use of syntactic distances for explaining mutual intelligibility. It would also be useful to validate our distance measurements for use in other disciplines, for example, for authorship attribution, forensic linguistics, stylometry, and studies of language contact and change.

Acknowledgments

We are grateful to Peter Kleiweg, whose RuG/L04 package was used to create the, dendrograms, and MDS plots shown in this article. We also thank the

anonymous reviewers for their valuable remarks on an earlier version of this article.

Funding

This research was supported by the Netherlands Organisation for Scientific Research grant number 360-70-430.

References

Baugh, A. and Cable, Th. (1978). *History of the English Language*, 3rd edn. Englewood Cliffs, NJ: Prentice Hall.

Brontë, A. (1847). *Agnes Grey*. London: T.C. Newby. Published under pseudonym Acton Bell.

Brontë, A. (1848). *The Tenant of Wildfell Hall*. London: T.C. Newby. Published under pseudonym Acton Bell.

Brontë, C. (1853). *Villette*. London: Smith, Elder & Company. Published under pseudonym Currer Bell.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**(3): 297–334.

Emonds, J. and Faarlund, J. T. (2014). *English: The Language of the Vikings*. Olomouc Modern Language Monographs Volume **3**. Olomouc: Palacký University.

Golubović, J. (2016). *Mutual Intelligibility in the Slavic Language Area*. Doctoral dissertation. Groningen: University of Groningen.

Grieve, J. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Doctoral dissertation. Flagstaff, AZ: Northern Arizona University.

Grieve, J. (2016). *Regional Variation in Written American English*. Studies in English Language. Cambridge: Cambridge University Press.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Doctoral dissertation. Groningen: University of Groningen.

Heeringa, W., Golubovic, J., Gooskens, C., Schüppert, A., Swarte, F., and Voigt, S. (2013). Lexical and orthograhic distances between Germanic, Romance and Slavic languages and their relationship to geographic distance. In Gooskens, C. and van Bezooijen, R. (eds), *Phonetics in Europe. Perception and Production*. Frankfurt am Main: Peter Lang, pp. 99–137.

Heeringa, W., Swarte, F., Schüppert, A., and Gooskens, C. (2014). Modeling intelligibility of written Germanic languages: Do we need to distinguish between orthographic stem and affix variation? *Journal of Germanic Linguistics*, **26**(4): 361–94.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts, *Literary & Linguistic Computing*, **22**(4): 405–19.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice Hall.

Kortmann, B. (2016). The Viking Hypothesis from a Dialectologist's Perspective. *Language Dynamics and Change* **6**(1), 27–30.

Kruskal, J. B. (1999). An overview of sequence comparison. In Sankoff, D. and Kruskal, J. (eds), *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, 2nd edn. Stanford: Center for the Study of Language and Information, pp. 1–44. 1st edition appeared in 1983.

Lauttamus, T., Nerbonne, J., and Wiersma, W. (2007). Detecting syntactic contamination in emigrants. The English of Finnish Australians. *SKY Journal of Linguistics*, **21**: 273–307.

Lebart, L. and Salem, A. (1994). *Statistique Textuelle*. Paris: Dunod.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, **10**(8): 707–10.

MacMurray, E. and Leenhardt, M. (2011). Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web. *Proceedings of the ICAI Conference 2011*, Las Vegas, pp. 605–11.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**(2): 209–20.

Miller, D. G. (2012). *External Influences on English: From Its Beginnings to the Renaissance*. New York: Oxford University Press.

Nelson, G., Wallis, S., and Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Nerbonne, J., Lauttamus, T., Wiersma, W., and Opas-Hänninen, L. L. (2010). Applying language technology to detect shift effects. In Norde, M., de Jonge, B., and Hasselblatt, C. (eds), *Language Contact: New Perspectives*. Amsterdam: Benjamins. Series IMPACT: Studies in Language and Society, pp. 27–44.

Nerbonne, J. and Wiersma, W. (2006). A measure of aggregate syntactic distance. In Nerbonne, J. and Hinrichs, E. (eds), *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, July, pp. 82–90.

Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill,

Sammon, J.W., Jr. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, **18**: 401–409.

Sampson, G. (2000). A proposal for improving the measurement of parse accuracy. *International Journal of Corpus Linguistics*, **18**(1): 53–68.

Sanders, N. C. (2007). Measuring syntactic difference in British English. In *Proceedings of the ACL 2007 Student Research Workshop*. Madison: Omnipress, pp. 1–6.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, **11**: 33–40.

Spruit, M. (2008). *Quantitative Perspectives on Syntactic Variation in Dutch Dialects*. Doctoral dissertation, Amsterdam: University of Amsterdam.

Swarte, F. (2016). *Predicting the Mutual Intelligibility of Germanic languages from linguistic and extra-linguistic factors*. Doctoral dissertation. Groningen: University of Groningen.

Szmrecsanyi, B. (2008). Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* **2**(1-2): 279–96. Special issue on "Language Variation", ed. by J. Nerbonne, C. Gooskens, S. Kürschner and R. van Bezooijen.

Torgerson, W. S. (1952). Multidimensional scaling: I Theory and method. *Psychometrika*, **7**: 401–19.

Torgerson, W. S. (1958). *Theory & Methods of Scaling*. New York, NY: Wiley.

Trudgill, P. (2016). Norsified English or Anglicized Norse? *Language Dynamics and Change*, **6**(1): 46–48.

Tufféry, S. (2007). *Data mining et statistique décisionnelle: l'intelligence des données*. Paris: Editions Technip.

Uzuner, Ö., Katz, B., and Nahnsen, Th. (2005). Using syntactic information to identify plagiarism. In *Proceedings of The Second Workshop on Building Educational Applications Using NLP*, Ann Arbor, Michigan, USA, June 2005, pp. 37–44.

Van Bezooijen, R. and Gooskens, C. (2005). How easy is it for speakers of Dutch to understand spoken and written Frisian and Afrikaans, and why? In Doetjes, J. and Van de Weijer, J. (eds), *Linguistics in the Netherlands*, Amsterdam: John Benjamins Publishers, vol. **22**, pp. 13–24.

Van Gelderen, E. (2016). Split infinitives in early Middle English. *Language Dynamics and Change*, **6**(1): 18–20.

Van Kemenade, A. (2016). English: The extent of Viking impact remains open. *Language Dynamics and Change*, **6**(1): 24–6.

Wiersma, W., Nerbonne, J., and Lauttamus, T. (2010). Automatically extracting typical syntactic differences from Corpora. *Literary and Linguistic Computing*, **26**(1): 107–24.

Zulu, P. N., Botha, G., and Barnard, E. (2008). Orthographic measures of language distances between the official South African languages. *Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, **29**(1): 185–204.

Notes

1. See http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/coverage.htm
2. See http://www.examenglish.com/CEFR/cefr.php
3. This example is taken from our own database.
4. That is, when using the linear version.
5. For comparing correlation coefficients we used the function r.test from the package psych version 1.4.8.11 in R, developed by William Revelle.

Appendix

Table A1 Syntactic distances between Germanic languages measured with the movement measure

		Reader				
		Danish	**Dutch**	**English**	**German**	**Swedish**
Stimulus	Danish	0	6.076	0.894	7.182	0.136
	Dutch	7.167	0	9.788	3.273	7.394
	English	1.182	8.364	0	9.546	3.227
	German	6.727	0.591	8.455	0	8.212
	Swedish	0.273	5.742	1.470	6.439	0

Table A2 Syntactic distances between Germanic languages measured with the indel measure

		Reader				
		Danish	**Dutch**	**English**	**German**	**Swedish**
Stimulus	Danish	0	2.621	1.879	1.849	0.682
	Dutch	1.788	0	1.727	1.924	2.758
	English	2.121	2.606	0	2.864	2.712
	German	1.394	1.061	1.773	0	1.955
	Swedish	0.606	3.015	1.773	1.773	0

Table A3 Syntactic distances between Germanic languages measured with the trigram measure

		Reader				
		Danish	**Dutch**	**English**	**German**	**Swedish**
Stimulus	Danish	0	0.317	0.161	0.340	0.043
	Dutch	0.329	0	0.325	0.168	0.366
	English	0.191	0.334	0	0.405	0.203
	German	0.358	0.075	0.442	0	0.379
	Swedish	0.067	0.310	0.147	0.356	0